Assessing objective characterizations of phonetic convergence

Gérard Bailly & Amélie Martin

GIPSA-Lab, Speech & Cognition Dpt, Grenoble - France
gerard.bailly@gipsa-lab.grenoble-inp.fr, amelie.martin@st.com

Abstract

This paper focuses on the study of the convergence between characteristics of speech segments– i.e. spectral characteristics of speech sounds – during live interactions between speaking dyads. The interaction data has been collected using an original verbal game called 'verbal dominoes' that provides a dense sampling of the acoustic spaces of the interlocutors. Two methods for characterizing phonetic convergence are here compared. The first one is based on a fine-grained analysis of the spectra of central frames of vowels (LDA) while the second one uses a more global speaker recognition technique (LLR). We show that convergence rates calculated by the two techniques correlate as the number of dominoes increases and that the LDA method well resists to the decrease of training and test material. We finally comment the impact of several factors on the computed convergence rates, i.e. interlocutors' familiarity and sex pairs.

Index Terms: phonetic convergence; dominos games; speaker recognition; interlocutors' familiarity.

1. Introduction

Giles et al [1-3] have introduced the Communication Adaptation Theory (CAT) that postulates that people in interaction will have the tendency to decrease the social distance between them by moving closer their behaviors (i.e. converge) or on the contrary accentuate their differences by moving apart (i.e. diverge). Researchers have notably examined adaptation of phonetic dimensions such as pitch [4], speech rate [5], loudness [6], dispersions of vocalic targets [7] as well as more global alignment such as rhythm at turn-taking [8]. The underlying assumption of CAT is that these social signals [also quoted as honest signals by 9] resulting from the adaptive behaviors of people in interaction are consciously or unconsciously processed and perceived to influence cognitive processes and production of verbal and co-verbal behavior of the interacting partners.

We assess here two different objective measures of phonetic convergence on data collected during a verbal dominos game played by 35 dyads with different sex and different previous mutual exposures. In particular we recorded interactions within members of two families.

2. Objective characterization of phonetic convergence

The influence of sensory input on speech production has also been investigated via the manipulation of certain characteristics of somatosensory feedback before or during speech production. Perturbations of jaw or lips movement [10], palate shape [11], pitch [12], formant frequencies [13] or spectral tilt [14] of the produced speech result in on-line, rapid and persistent (after-effect) compensations in speech production. These compensations are large and conservative: even when perturbations are very subtle [15], speakers and singers seem to rely on an internal model to regulate their productions that privileges planned sensory objectives.

In the later case, the objective characterization of the adaptive behavior is often straightforward: known perturbations are applied to crucial characteristics for a given speech production task – i.e. articulatory variable, vocal tract geometry or acoustic feature – and researchers mainly focus on compensatory effects on the perturbed feature. In the case of perturbations induced by the environment – i.e. ambient or interactive speech – with no explicit and controlled manipulation of pre-recorded speech, the space of free variables is much larger and the quest for an objective characterization of the adaptive behavior is much more challenging.

2.1. Phonetic cues

The most popular approach consists in exploring a set of features that mirror the expected sensory-motor adaptation induced by the particular experimental design. Several reference works have notably examined the adaptation of specific phonetic contrasts [formants and durations of specific sounds in 7, voice onset times (VOT) in 16, 17] via cross language/dialectal studies involving productions of monolinguals or bilinguals as a function of ambient language. When not focusing on dialectal variations or selected features, obtaining a robust and global objective estimation of the amplitude of adaptation that could be confronted to subjective ratings is still an open issue.

2.2. Holistic characterization

Researchers have proposed methods that provide holistic measurements of phonetic accommodation based on a global comparison between the temporal and spectral characteristics of two sets of speech signals. One of the first key study was performed by Delvaux and Soquet [7]. They compared the global spectral characteristics of target sounds thanks to a linear discriminant analysis (LDA) between speech parameters produced by speakers of the different dialects during pre-test, interactive and post-test sessions. LDA was in fact used to select the most discriminative dimensions among the set of 20 Mel-frequency cepstral coefficients (MFCC) that represent the general distribution of signal energy along the frequency axis up to 10500 Hz. Aubanel and Nguyen [18] tested different levels of convergence in the dyads (towards the interlocutor, the interlocutor's group or accent) using LDA performed on spectral characteristics of specific segments.

Kim et al [19, 20] studied accommodation of words pronounced before vs. after a phonetic accommodation session where they listened to native American-English vs. nonnative English uttered by Korean speakers. They computed a global similarity cost between words using dynamic time warping (DTW). Kim et al used also MFCC as the characterization of the spectral slices. They compared convergences rates computed by normalized DTW cumulated distances vs. results of XAB tests and found that the perceived accommodation patterns could be partially predicted by the DTW distance

changes. Pardo [21] also found similar results. Although DTW seems to provide reliable holistic measurements of phonetic accommodation, it has some important limitations: (a) Speakers should pronounce identical words; (b) Several references of each word are to be aligned with accommodated productions so that to mirror intra-speaker variability. More sophisticated word models may be built using statistical models such as HMM that could be aligned with Figure 1thanks a Viterbi algorithm (analog to DTW) but their training require much more data than is often available in accommodation experiments; (c) It is rather difficult to sort out contributions of prosodic, phonological, allophonic and phonetic variations to the computed cumulated distance.
In the following we will thus use the method proposed by Delvaux and Soquet [7] – i.e. LDA performed on MFCC of target frames of a set of given allophones – as the holistic measurement of phonetic accommodation. Special care has been given to the labeling of allophonic variation (so that to compare phonetic accommodation within identical phonological spaces). We also validate the speakers' models by distinguishing between training and validation data.

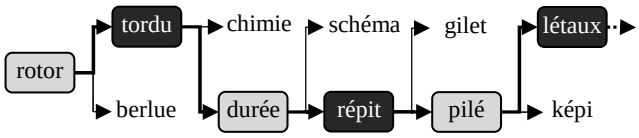

Figure 1. *The 6 first speech dominos of the game. Except for the first display of the initiator, both speakers are presented with a pair of real and frequent words. They have to read aloud the one that starts with the last syllable of the word just uttered previously by their interlocutor. Correct words are here circled and linked by dark lines. Speakers choose in alternation between two written words. The correct path is unique but not predictable.*

3. Data & experiments

Most experimental paradigms involve repetitions of speech units – isolated sounds (vowels, syllables, words, whole sentences [see 20] – either explicitly with reading or shadowing tasks or implicitly by asking interlocutors to refer to items of the common ground, such as landmarks in the map task [22] or elements of scenes in the diapix [23]. Post-test sessions can then be used to sort out effects of immediate (stimuli-dependent) imitation from mimesis (i.e. deeper changes of sensory-motor representations of units). The experimental paradigms used so far either collect few instances [a dozen in 18, four key phrases in 24, four key phrases also in 25] of few key segments or many instances of a very small set of key segments (two in Delvaux and Soquet). For several studies, the segments are also chosen to maximize dialectal variation: this choice is questionable since it remains to be shown that subjects effectively negotiate these critical segments at first, before or more easily than others. Since convergence is rather segment-dependent, it is interesting to study the impact of speakers' alignment more holistically on their entire phonetic repertoire. Babel [26, 27] notably studied the impact of social information on imitation in vowels. She asked speakers to shadow 50 low-frequency words uttered by two talkers speaking Californian English in 6 different conditions that implicitly influenced convergence patterns expressed as relative distances between target first and second formants.

3.1. Experimental paradigm: verbal dominos

Our study is based on the reading aloud several hundreds of mid- to high- frequency words that maximally covers the phonetic repertoire of the target language, here French.
For our experiments, we developed an interaction paradigm called "Verbal Dominoes" [28], where speakers alternatively choose and utter words that begin with the same syllable as the one ending the previous word. Such rhyme games are part of the children's folklore, played in playgrounds [29] and widely used in primary school, for example for language learning. While "Word chain" - also known as Grab on Behind, Last and First, and Alpha and Omega – consists in coming up with words that begin with the letter or letters that the previous word ended with, we chain here spoken words. This verbal game is also known in Japon as *Shiritori* that consists in chaining kana syllables. The rule of the game is quite simple. Speakers are presented with a pre-selection of written words and have to choose the one that begins with the same syllable as the final syllable of the word previously uttered by the interlocutor (see Figure 1).
We selected here words with mid- to high-lexical frequencies so that to uniformly collect allophonic variations of the eight peripheral oral vowels of French: [a], [ɛ], [e], [i], [y], [u], [o], [ɔ]. Alternatives are here limited to two dissyllabic words in order to limit the cognitive load and ease the running of successive sessions.
We established two chains of dominoes that collect respectively around 20 vs. 40 exemplars of each peripheral oral vowel. The first chain is referenced as the *baseline* chain. It chains 183 words. We extended this chain by appending 165 dominos. This will be named the *extended* chain. Note that both chains begin with the same 183 words.

4. Speakers and conditions

Overall, convergence rates are often significant but weak and strongly depend on the dyads. A strong implicit assumption made by most studies is the hypothesis that convergence could be rapid and observed within the few minutes of laboratory experiments. Most studies in fact confronts speakers unknown to each other: Our first dyads) also consisted of unknowns. The measured convergence rates were quite small. In order to observe a large variety of convergence patterns, we then explore the convergence patterns between people with prior mutual exposure: friends [see also 25] and family members. In contrast with the long-term investigation between roommates conducted by Pardo, Gibbons et al. [25] who did not find large convergence rates (but using non interactive speech), our data show that a long-term exposure together with positive social links indeed result in larger convergence rates.
The speakers pronounced dominoes under different conditions. The acoustic references for each speaker are first collected during a *pre-test*. During this condition, they read aloud in isolation all words that will be pronounced by the two speakers during the dominoes' game. They are presented in random order to both participants, sitting alone in the same quiet environment. Once each speaker has performed the pre-test alone, they are introduced to each other and the verbal game is performed.
In this paper, we the pre-test condition and the interactive game played during the four experiments:

- **Experiment I (12 dyads)**: speakers sit in two different rooms and communicated through close microphones and

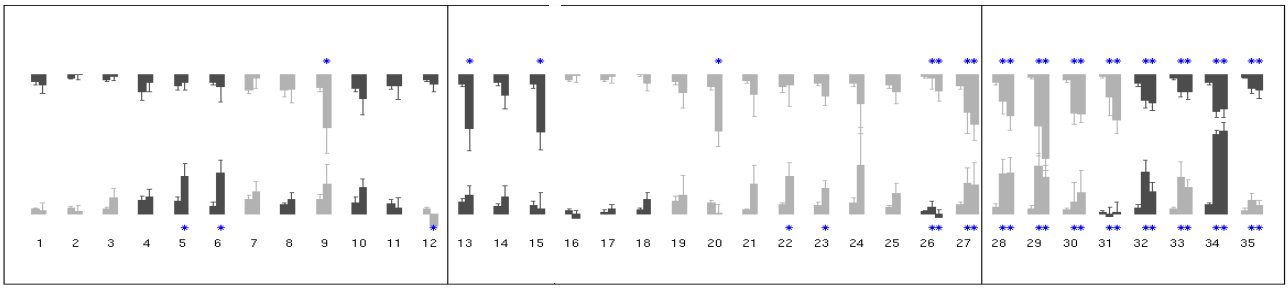

Figure 2. *Convergence rates computed by LDA for the 35 dyads grouped by conditions. From left to right: with the baseline set of dominos, (a) unknowns; with the extended set of dominos, (b) friends and (c) family members.*

headphones. They exchanged the *baseline* chain. Speakers were unknown to each other.

- **Experiment II (13 dyads)**: speakers sit face-to-face with two screens back-to-back displaying the alternative words. Eye contact was possible. They exchanged the *extended* chain. We studied here dyads of good friends (mean relation of 2 years ± 9 months from 6 months to 28 years).
- **Experiment IV (10 dyads)**: same as Experiment II but between members of two families (mean of 30 years ± 5 months from 19 years to 53 years).

Subjects are instructed to avoid speech overlaps and repairs in order to ease automatic segmentation and alignment. Recordings are performed using Sony Tie-Pin *ECM-C115* microphones with batteries.

4.1. Objective characterizations

For each dyad, a statistical model is built for each speaker using pre-test data. Part of this data is used to effectively train the model and the other part (namely validation data) is used to quantify the robustness and possible overfitting of the model. We also verify that our speakers' models are not sensitive to intra-speaker variability such as provided by simple repetitions of stimuli [see similar concern in 20]. Normalized distances (or log-probability) between the test data and the models of each dyads are then used to estimate the degree of phonetic convergence between speakers' productions. Since speakers read aloud dominos that will be uttered by both interlocutors, the amount of pre-test material equals to the double of the amount of test data. Training, validation and test materials are thus of equal sizes.

The split between training and validation is performed 20 times. All results presented in section 4.2 exploit statistics of these multiple simulations.

Pre-processing

All models below use spectral representations of speech frames. Cepstral Mean Subtraction [30] is applied after the computation of Mel-Frequency Cepstral Coefficients for each set of data to reduce residual mismatch between microphones and recording environments.

Linear discriminant analysis of vocalic spectra

A linear discriminant analysis (LDA) of MFCC target vocalic spectra, similar to what was proposed by Delvaux and Soquet [7], was first performed [31]. The principle is quite simple: for each vowel, LDA determines the acoustic space in which the productions of one speaker during his pre-test differs the most from the ones of his interlocutor. Note that each vocalic nucleus is labeled with the proper phonetic category: we focus here on how the pronunciation of each intended vowel is influenced by the conditions. The MFCC of the central frames of the validation and test material are then projected on the first discriminant axis separating speaker-specific spaces for each corresponding vowel. This scalar projection is named $d_{LDA}()$ in the following. We calculate convergence rates $C_{LDA}(s1,s2,v)$ of speaker s1 towards speaker s2 for each vowel *v* by normalizing the mean distance between the projections of test data of s2 during interaction with s1 and those of the training data of s1 by the distance between projections uttered during the pre-test:

$$C_{LDA}(s1,s2,v) = \frac{mean(d_{LDA}(P_{s1}) - d_{LDA}(I_{s1s2}))}{mean(d_{LDA}(P_{s1}) - d_{LDA}(P_{s2}))} \quad (1)$$

Speaker recognition techniques

We compare here results obtained using the LDA method described above with a second method based on speaker recognition techniques [32]. We used the Alizee platform [33]. Acoustic spaces of speakers are modeled by Gaussian mixtures models (GMM), one of the most popular techniques for text-independent speaker recognition [34]. The speaker decision task mainly consists in a basic statistical test between two hypotheses: (1) H_S: the speech characteristics *y* has been produced by the hypothesized speaker S and (2) $H_{\neg S}$: *y* is not from the hypothesized speaker S (often called the model of the "world"). In our case, H_S and $H_{\neg S}$ are the models of the two speakers of the dyad: the "world" $\neg S$, usually trained with speech samples from a large number of speakers representative of the speaker population as a whole, corresponds only to the interlocutor's model. We then compute the log-likelihood of samples Y to have been produced by speaker s1 but not by speaker s2:

$$LLR_{s1s2}(Y) = \sum_{y \in Y} \log\left(\frac{p(y \mid H_{s1})}{p(y \mid H_{s2})}\right) \quad (2)$$

GMMs here have M=64 components and the *y* components are MFCC coefficients computed every 10ms.

These GMMs are trained in order to maximize $LLR_{s1s2}(P_{s1})+LLR_{s2s1}(P_{s2})$ over the set of training frames P_{s1} and P_{s2} uttered respectively by speakers *s1* and *s2* during the pre-test. This sum corresponds to the global distance between acoustic spaces of the two speakers.

The convergence rate of *s1* "towards" *s2*, called $C_{LLR}(s1,s2)$ is then taken as the relative quotient between the difference of a speaker's LLR (here *s1*) calculated with his own model on frames P_{s1} and during interaction (I_{s1s2}) and the difference of LLR calculated with the two interlocutor's model on the pre-test (P_{s1}).

$$C_{LLR}(s1,s2) = \frac{LLR_{s1s2}(P_{s1}) - LLR_{s1s2}(I_{s1s2})}{LLR_{s1s2}(P_{s1}) - LLR_{s1s2}(P_{s2})} \quad (3)$$

where I_{s1s2} is the set of frames uttered by speaker *s1* when interacting with speaker *s2*. So, if we don't have any convergence, $I_{s1s2}=P_{s1}$ and $C_{LLR}(s1, s2) = 0$.

Once again, the calculation C_{LLR} is repeated 20 times (random split between training and validation) and the mean and standard deviation of these convergence rates for the validation and the test data are computed.

Note that the speaker recognition technique calculates a global convergence rate (e.g. not only on specific vocalic targets) without any a priori segmentation. It also skips the problem of assigning precise phonetic labels or features such as palatalization or devoicing, accounting for particular dialectal variations or idiosyncrasies.

GMMs are here trained only on speech frames, i.e. any silence exceeding 300ms is discarded from training, validation and test data. The average available durations of these non-silent frames are respectively 45.7±5.2s for the baseline corpus and 87±7s for the extended corpus. With an analysis rate of 10ms, 4565 vs. 8700 frames on average are used to train, validate and test the GMM. This should be compared to the 20 vs. 40 target frames per vowel used for LDA.

4.2. Results

The Figure 2 shows the results obtained with the linear discriminant analysis on the 35 interactions. We have inverted results for one interlocutor (top line) to better illustrate the convergence between both subjects for each dyad. For each dyad, the bar on the left corresponds to our validation condition (e.g. convergence rates computed on one half of the pre-test), the second bar illustrates the convergence rate computed for the interaction. Convergence rates we obtained range from -0.05 to 0.6. This large range was obtained thanks our variety of prior exposure between dyads.

We compared the distributions of convergence rates obtained by LDA and LLR on the extended corpus (considering only experiments II and III) using different sizes of training and test material. Figure 3 shows that LDA seems to be less sensitive to corpus size, although only vocalic targets are considered.

The convergence rates were submitted to a repeated measures ANOVA to test for the effects of *method* (LDA vs LLR), *session* (pre-task versus interaction), *familiarity* (unknown, friend vs. family), *sex of the subject* and *sex of the interlocutor* (female vs males. Since the first half of the extended chain is strictly the baseline one, we performed the analysis on the convergence rates obtained using the first 183 dominoes. Results confirm the main effect of *session* [$F=1355$, $p<10^{-16}$], *familiarity* [$F=160$, $p<10^{-16}$], *sex of the speaker* [$F=58$, $p<10^{-14}$] and *sex of the interlocutor* [$F=19$, $p<10^{-5}$]. The factor *method* is not significant [$F=2.64$, $F>0.1$]. We aggregated LDA & LLR estimations and further explored the convergence rates during interaction. We observe a highly significant interaction between *sexes* of the dyad [$F=289$, $p<10^{-16}$] and a three-ways interaction between *familiarity* and *sexes* [$F=14.8$, $p<10^{-7}$]: same sex dyads converge more and opposite sex and this convergence is amplified for familiar dyads.

We indeed obtained larger convergence rate for pairs of subject from the same family in comparison with unknowns and friends, particularly for two pairs corresponding to interactions between sisters (pair 29) or brothers (pair 34). It is interesting to notice that, for the first family (pairs 28 to 31), convergence mirrors social hierarchy [35]: in fact, parents (respondents of pairs 30 and 31) do not significantly change their behavior between the pre-test and the interaction. We do not see this phenomenon for the second family (pairs 32 to 35). While the convergence is larger for the pair of brothers (pair 34), convergence is modest between the son and his parents (pairs 32 & 33) and even weaker for the pair composed of the brother and the sister (pair 35). This tendency confirms that convergences is larger between interlocutors with equivalent social status, for same-sex pairs and particularly for females (see Figure 4).

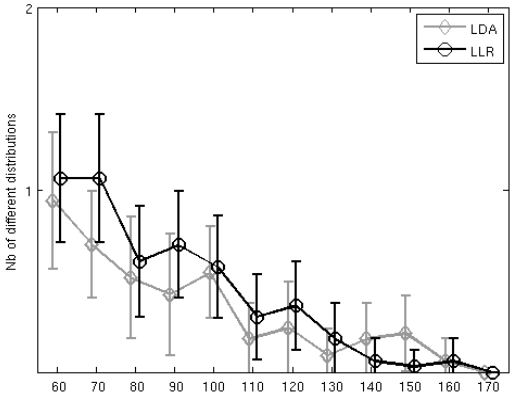

Figure 3: *Convergence rates: average number of distributions computed over various numbers of dominos that are statistically different from the one computed with the full set. As expected, LLR is rather stable with large training sizes (>130 dominoes) but degrades estimation when fewer examples are available.*

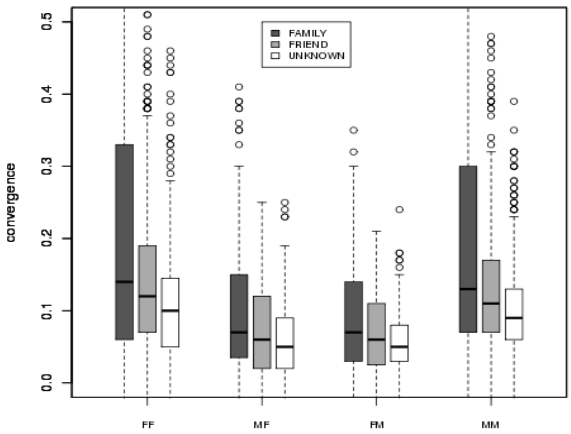

Figure 4: Mean convergence rates averaged over LDA and LLR as a function of familiarity and sex pairs.

5. Conclusions

We compared two different methods for characterizing phonetic convergence using stimuli gathered during four sets of experiments involving 35 French dyads. The speaker recognition technique provides rather consistent results in comparison with detailed phonetic analysis focusing on vocalic segments once sufficient samplings of the speakers' acoustic spaces are made available. This technology opens the way for analyzing more complex conversational situations, notably to observe if our goal-directed task influences the ecological validity of the results.

We observed almost no divergence but found several occurrences of strong and significant phonetic convergence depending on dyads, sex of pairs and also social relationships. The strongest convergence rates were observed for same-sex pairs with well-established social relationships.

6. References

[1] Giles, H. and R. Clair, *Language and Social Psychology*. 1979, Oxford: Blackwell.

[2] Giles, H., et al., *Speech accommodation theory: The first decade and beyond*, in *Communication Yearbook*, M.L. McLaughlin, Editor. 1987, Sage Publishers: London, UK. p. 13-48.

[3] Giles, H., J. Coupland, and N. Coupland, *Contexts of Accommodation: Developments*. Applied Sociolinguistics. 1991, Cambridge: Cambridge University Press. 332.

[4] Gregory, S.W., S. Webster, and G. Huang, *Voice pitch and amplitude convergence as a metric of quality in dyadic interviews.* Language and Communication, 1993. **13**.

[5] Edlund, J., M. Heldner, and J. Hirschberg. *Pause and gap length in face-to-face interaction*. in *Interspeech*. 2009. Brighton.

[6] Kousidis, S., et al. *Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues*. in *Interspeech*. 2008. Brisbane.

[7] Delvaux, V. and A. Soquet, *The influence of ambient speech on adult speech productions through unintentional imitation.* Phonetica, 2007. **64**: p. 145-173.

[8] Benus, S. *Are we 'in sync': Turn-taking in collaborative dialogues*. in *Interspeech*. 2009. Brighton, UK.

[9] Pentland, A., *Honest Signals: how they shape our world*. 2010, Cambridge, MA: MIT Press.

[10] Bauer, A., L. Jäncke, and K.T. Kalveram, *Mechanical perturbation of jaw movements during speech: effects on articulation and phonation.* Perceptual and Motor Skills, 1995. **80**(3 Pt 2): p. 1108-1112.

[11] Honda, M., A. Fujino, and A. Kaburagi, *Compensatory responses of articulators to unexpected perturbations of the palate shape.* Journal of Phonetics, 2002. **30**: p. 281-302.

[12] Jones, J.A. and K.G. Munhall, *Perceptual calibration of F0 production: Evidence from feedback perturbation.* Journal of the Acoustical Society of America, 2000. **108**: p. 1246-1251.

[13] Houde, J.F. and M.I. Jordan, *Sensorimotor adaptation in speech production.* Science, 1998. **279**: p. 1213-1216.

[14] Shiller, D.M., et al., *Perceptual recalibration of speech sounds following speech motor learning.* Journal of Acoustical Society of America, 2009. **125**(2): p. 1103–1113.

[15] Keough, D. and J.A. Jones, *The sensitivity of auditory-motor representations to subtle changes in auditory feedback while singing.* Journal of the Acoustic Society of America, 2009. **126**(2): p. 837-846.

[16] Fowler, C.A., et al., *Cross language phonetic influences on the speech of French-English bilinguals.* Journal of Phonetics, 2008. **36**(4): p. 649-663.

[17] Sancier, M. and C.A. Fowler, *Gestural drift in a bilingual speaker of Brazilian Portuguese and English.* Journal of Phonetics, 1997. **25**: p. 421-436.

[18] Aubanel, V. and N. Nguyen, *Automatic recognition of regional phonological variation in conversational interaction.* Speech Communication, 2010. **52**: p. 577-586.

[19] Kim, M., W. Horton, and A. Bradlow, *Phonetic convergence in spontaneous conversations as a function of interlocutor language distance.* Journal of Laboratory Phonology, 2011. **2**: p. 125-156.

[20] Kim, M., *Phonetic accommodation after auditory exposure to native and nonnative speech*. 2012, Northwestern University: Evanston, IL. p. 153.

[21] Pardo, J.S., *Expressing oneself in conversational interaction*, in *Expressing Oneself/Expressing One's self: Communication, cognition, language, and identity*, E. Morsella, Editor. 2010, Psychology Press: London. p. 183-196.

[22] Anderson, A.H., et al., *The HCRC Map Task corpus.* Language and Speech, 1991. **34**: p. 351-366.

[23] Van Engen, K.J., et al., *The Wildcat Corpus of native- and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles.* Language & Speech, 2010. **53**: p. 510-540.

[24] Pardo, J.S., *On phonetic convergence during conversational interaction.* Journal of the Acoustical Association of America, 2006. **119**(4): p. 2382–2393.

[25] Pardo, J.S., et al., *Phonetic convergence in college roommates.* Journal of Phonetics, 2012. **40**(1): p. 190-197.

[26] Babel, M., *The effect of talker image on phonetic convergence.* Journal of Acoustical Society of America, 2008. **124**: p. 2559.

[27] Babel, M.E., *Phonetic and social selectivity in speech accommodation*, in *Department of Linguistics*. 2009, University of California: Berkeley, CA. p. 181.

[28] Bailly, G. and A. Lelong. *Speech dominoes and phonetic convergence*. in *Interspeech*. 2010. Tokyo.

[29] Arléo, A., *Un jeu de dominos verbal: Trois p'tits chats, chapeau d'paille*, in *Chants enfantins d'Europe*, A. Arléo, et al., Editors. 1997, L'Harmattan: Paris. p. 33-68.

[30] Furui, S., *Cepstral analysis technique for automatic speaker verification.* IEEE Transactions on Acoustics, Speech and Signal Processing, 1981. **29**: p. 254-272.

[31] Lelong, A. and G. Bailly, *Study of the phenomenon of phonetic convergence thanks to speech dominoes*, in *Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issue*, A. Esposito, et al., Editors. 2011, Springer Verlag: Berlin. p. 280-293.

[32] Lelong, A. and G. Bailly. *Characterising phonetic convergence with speaker recognition techniques.* in *The Listening Talker Workshop*. 2012. Edinburgh.

[33] Charton, E., et al. *Mistral: an open source biometric platform*. in *25th Symposium on Applied Computing (SAC)*. 2010. Sierre, Switzerland.

[34] Reynolds, D., *Speaker identification and verification using Gaussian mixture speaker models.* Speech Communication, 1995. **17**(1): p. 91-108.

[35] Campbell, N. *Listening between the lines; a study of paralinguistic information carried by tone-of-voice*. in *International Symposium on Tonal Aspects of Languages: Emphasis on Tone Languages*. 2004. Beijing.